Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations

Year: 2023 | Citations: 287 | Authors: Josh A. Goldstein, Girish Sastry, Micah Musser, Renée DiResta, M. Gentzel

---

Abstract

Generative language models have improved drastically, and can now produce realistic text outputs that are difficult to distinguish from human-written content. For malicious actors, these language models bring the promise of automating the creation of convincing and misleading text for use in influence operations. This report assesses how language models might change influence operations in the future, and what steps can be taken to mitigate this threat. We lay out possible changes to the actors, behaviors, and content of online influence operations, and provide a framework for stages of the language model-to-influence operations pipeline that mitigations could target (model construction, model access, content dissemination, and belief formation). While no reasonable mitigation can be expected to fully prevent the threat of AI-enabled influence operations, a combination of multiple mitigations may make an important difference.